14

Increasing Computational Efficiency

14.1 Seven Techniques

With today's high-speed computers, drawing large numbers of subsamples with replacement (the bootstrap) or without (the permutation test) is no longer a problem, unless or until the entire world begins computing resampling tests at one time! To prepare for this eventuality, and because computational efficiency is essential in the search for more powerful tests, a primary focus of research in resampling today is the development of algorithms for rapid computation.

There are seven main computational approaches, several of which may be and usually are employed in tandem, as follows:

1. The *Monte Carlo*, in which a sample of the possible rearrangements is drawn at random and these samples are used in place of the complete permutation distribution.
2. *Rapid enumeration and selection algorithms*, whose object is to provide a rapid transition from one rearrangement to the next.
3. *Recursive relationships*, which reduce the number of computations.
4. *Branch and bound algorithms* that eliminate the need to evaluate each individual rearrangement.
5. *Gibbs sampling.*
6. Solution through *characteristic functions and fast Fourier transforms.*
7. *Asymptotic approximations*, for use with sufficiently large samples.

In the following sections, we consider each of these approaches in turn.

14.2 Monte Carlo

Instead of examining all possible rearrangements, we can substantially reduce the computations required by examining only a small but representative random sample [Dwass, 1957; Barnard, 1963]. In this process, termed a Monte

Carlo, we proceed in stages: 1) We rearrange the data at random; 2) we compute the test statistic for the rearranged data and compare its value with that of the statistic for the original sample; and 3) we apply a stopping rule to determine whether we should continue sampling, or whether we are already in a position to accept or reject.

The program fragments reproduced in Chapters 3 through 5 of this text use the Monte Carlo approach. In the not necessarily optimal computer algorithm introduced in those chapters, all observations in all subsamples are loaded into a single linear vector $\boldsymbol{X} = \{\boldsymbol{X}[0], \boldsymbol{X}[1], \ldots, \boldsymbol{X}[N-1]\}$. Then, a random number is chosen from the set of integers $0, 1, \ldots, I$ with $I = N-1$ initially. If the number we choose is i, $\boldsymbol{X}[i]$ is swapped with $\boldsymbol{X}[I]$ in a three-step process:

```
temp := X[i];
X[i] := X[i];
X[N-1] := temp;
```

and I is decremented. This process is repeated until we have rearranged the desired number of observations and are ready to compute the test statistic for the new rearrangement.

We dont't always need to reselect all N observations. For example, in a two-sample comparison of means, with $N = n + m$, our test statistic only makes use of the last m observations. Consequently, we only need to choose m random numbers each time.

After we obtain the new value of the test statistic, we compare it with the value obtained for the original data. We continue until we have examined N random rearrangements and N values of the test statistic. Typically, N is assigned a value between 100 and 1600, depending on the precision that is desired (see Section 14.2.2 and Marriott [1979]). Through the use of a Monte Carlo, even the most complicated multivariate experimental design can be analyzed in les than a minute on a desktop computer.

14.2.1 Stopping Rules

If a simple accept/reject decision is required, we needn't perform all N calculations, but can stop as soon as it is obvious that we must accept or reject the hypothesis at a specific level. In practice, we use a one-sided stopping rule based on the 10% level. Suppose in the first n rearrangements we observe a fraction H_n with a value of the test statistic that is as, or more, extreme than the value for the original observations. If $H_n > 0.1N$, then we accept the hypothesis at the 10% level. Otherwise, we continue until $n = N$ and report the exact percentage of rejections. Besag and Clifford [1991] and Lock [1991] describe two-sided sequential procedures in which the decision to accept, reject, or continue is made after each rearrangement is examined.